An Algorithm for Sequence Searching of Peptide Spectra Generated via Electron Transfer Dissociation

Lewis Y. Geer[1]; Dina L. Bai[2]; Jeffrey Shabanowitz[2]; Joshua J. Coon[2]; John E.P. Syka[2]; An Chi[2]; Beatrix M. Ueberheide[2]; Jeffrey A. Kowalak[3]; Sanford P. Markey[3]; Stephen H. Bryant[1]; Donald F. Hunt[2]
[1]National Library of Medicine, NIH, Bethesda, MD; [2]University of Virginia, Charlottesville, VA; [3]National Institute of Mental Health, NIH, Bethesda, MD

Overview and Introduction

- Electron Transfer Dissociation [ETD](1) is a novel method for MS/MS peptide sequence analysis particularly suited for characterization of post translational modifications [PTMs].
- While ETD spectra of peptides can be identified by sequence search algorithms developed for Collisionally Activated Dissociation [CAD] peptide spectra, existing algorithms can be computationally intensive and yield false positives that require manual validation, in part due to the large number of possible charge states per peptide.
- Recently, ETD has been used to develop a new method for the identification of intact proteins by creating spectra containing a series of c-ions from the amino-terminus of the protein and z-ions from the carboxy-terminus (2).
- Presently, there are no sequence search algorithms that search ETD spectra of intact proteins.
- The large number of possible modifications that can be present in some peptides and proteins also pose a challenge in statistical significance and computational efficiency.
- We have extended the Open Mass Spectrometry Search Algorithm [OMSSA](3) to efficiently search ETD spectra of modified peptides and proteins with increased sensitivity and specificity.

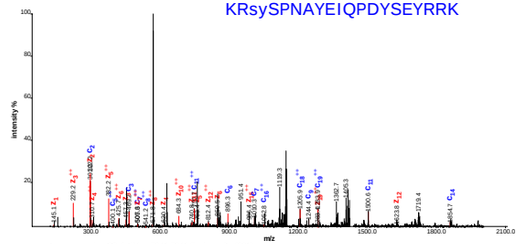
Methods

- The OMSSA statistical scoring model was developed using classical hypothesis testing where the null hypothesis is that the match between a given spectra and a sequence is entirely random.
- The OMSSA algorithm is written in C++ using the NCBI C++ and C toolkits. These cross-platform libraries allow OMSSA to access memory mapped NCBI BLAST sequence libraries.
- Spectra were obtained from a commercial Quadrupole Linear Ion Trap, the Finnigan LTQ mass spectrometer (Thermo Electron, Waltham, MA) equipped with a modified nanoflow electrospray ionization source or a nanospray robot. The LTQ was modified to accommodate a Finnigan 4500 CI source (Thermo Electron) placed at the rear of the instrument to facilitate ETD.

Results

- OMSSA was modified to search for the charge +1 and +2 c- and z-product ions generated by ETD and to search for the highly charged precursor ions that can be present in ETD.
- The following matched spectrum has a precursor charge of +5. The match is to yeast protein Yfr003cp with phosphorylation of S:3 and phosphorylation of Y:4. The E-value of the hit is 2.018 x10-304. High intensity unmatched peaks are charge reduced precursor ions.

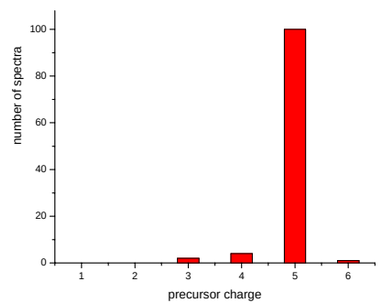
KRsySPNAYEIQPDYSEYRRK



spectrum name: LysC8_fatIMac.5041.5041.5_clean.dta (201)

Charge Independence

- ETD precursor ions can have a large range of experimentally unknown charge states. The following histogram enumerates charge states for significant hits to a yeast phosphopeptide set searched from precursor charge 3 to charge 6:



- The OMSSA scoring model accounts for precursor charge state. In the search described above, only one spectrum had hits below an E-value threshold of 0.1 with more than one charge state.

Modification Enumeration

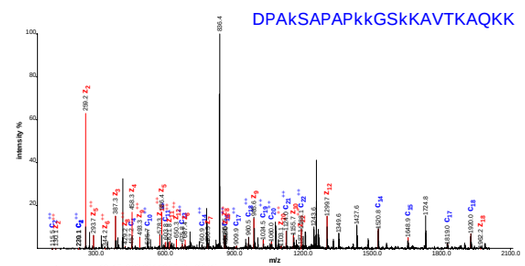
- Specification of variable modifications in an MS/MS sequence search algorithm can lead to a combinatorial explosion of theoretical peptide spectra for a given peptide, where each theoretical spectrum corresponds to a different combination of variable modifications applied to the peptide. This combinatorial explosion can adversely affect scoring and search speed.
- OMSSA was modified to allow specification of the maximum number of modification combinations to attempt per peptide. It was also modified to enumerate the modification combinations in order of least modified to most modified.
- To give an example, assume that the limit on the number modification combinations per peptide is 11 and that the theoretical peptide being searched is STYY with phosphorylation of S, T, and Y. The combinations that OMSSA will test are:

1	S	T	Y	Y
2	S	T	Y	Y
3	S	T	Y	Y
4	S	T	Y	Y
5	S	T	Y	Y
6	S	T	Y	Y
7	S	T	Y	Y
8	S	T	Y	Y
9	S	T	Y	Y
10	S	T	Y	Y
11	S	T	Y	Y

where a black character represents no modification and a highlighted red character represents a modification. OMSSA tries the combinations with the least number of variable modifications and then adds modifications until the maximum number of modification combinations is reached.

- An example match to a highly modified human histone H2B.2 peptide is shown below. The peptide was digested with Asp-N+Glu-C and has the modifications: di-methylation of K:4 and K:11, tri-methylation of K:10, and acetylation of K:14. The E-value of the hit was 4 x 10^{-304}. High intensity unmatched peaks are charge reduced precursor ions.

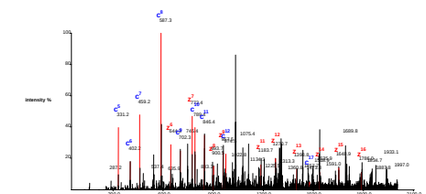
DPAkSAPAPkkGSkKAVTKAQKK



spectrum name: 12_15_04C.604.604.5.dta (1041)

Whole protein identification

- Coon et al. (2) recently described a novel method using ETD to sequence the amino and carboxy terminal ions of whole proteins. In one form, the method yields approximately 17 +1 c-ions from the amino terminus and 17 +1 z-ions from the carboxy terminus. OMSSA was modified to calculate these ions and to limit its scoring to the resulting set of product ions.
- Six spectra from two chicken histones, human histone, bovine albumin, bovine ubiquitin, and bovine cytochrome C were analyzed by searching the spectra against species specific subsets of the NCBI nr protein library. A example hit to a novel chicken histone H2B is shown below:



- All spectra found significant hits to true positives with the top hit being a true positive:

Protein	E-value of top hit
bovine cytochrome C	1.029E-5
bovine albumin	4.210E-08
bovine ubiquitin	4.84E-10
chicken H2A	1.382E-304
chicken hypothetical histone	1.250E-304
human histone H2B	3.969E-304

- All false positives above threshold were slightly modified versions of the true positives. This result is not unexpected as the OMSSA statistical model only estimates the probability of random matches.
- The search speed of the algorithm was 115,385 sequences per second per spectrum on one 3.06 GHz Intel Xeon processor.

Conclusions

- We have modified OMSSA to search ETD spectra of post translationally modified proteins and peptides in a specific and efficient manner.
- A public search service and downloadable executables are available at http://pubchem.ncbi.nlm.nih.gov/omssa.

References

(1) Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Proc Natl Acad Sci U S A. 2004 Jun 29;101(26):9528-33.
(2) Coon JJ, Ueberheide B, Syka JEP, Dryhurst DD, Ausio J, Shabanowitz J, Hunt DF. Proc Natl Acad Sci U S A, in press.
(3) Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. J Proteome Res. 2004 Sep-Oct;3(5):958-64.